ON MOVING AVERAGES AND ASYMPTOTIC EQUIPARTITON OF INFORMATION

by

R. NAIR

Department of Mathematical Sciences,
University of Liverpool,
Liverpool L69 3BX,
U.K.

Abstract : We prove a moving average version of the Shannon-McMillan-Breiman theorem.

AMS Subject Classification : Primary 28D05, 94A17; secondary 28A65, 28D20, 60F15

Keywords : Ergodic theory of information, asymptotic equipartion property, moving average ergodic theorem.

§1 Introduction :

We begin by introducing some notation. Let Z be a collection of points in $\mathbf{Z} \times \mathbf{N}$ and let

$$Z^h \ = \ \{(n,k) \ : \ (n,k) \ \in \ Z \text{ and } k \ \geq \ h\},$$

$$Z^h_\alpha \ = \ \{(z,s) \ \in \ \mathbf{Z}^2 \ : \ |z \ - \ y| \ < \ \alpha(s \ - \ r) \text{ for some } (y,r) \ \in Z^h\}$$

and

$$Z^h_\alpha(\lambda) \ = \ \{n : \ (n,\lambda) \ \in \ Z^h_\alpha\}. \qquad (\lambda \in \mathbf{N})$$

Geometrically we can think of Z^1_α as the lattice points contained in the union of all solid cones with aperture α and vertex contained in $Z^1 = Z$. We say a sequence of pairs of natural numbers $(n_l, k_l)_{l=1}^\infty$ is *Stoltz* if there exists a collection of points Z in $\mathbf{Z} \times \mathbf{N}$, and a function $h \ = \ h(t)$ tending to infinity with t such that $(n_l, k_l)_{l=t}^\infty \ \in \ Z^{h(t)}$ and there exist h_0, α_0 and $A \ > \ 0$ such that for all integers $\lambda \ > \ 0$ we have $|Z^{h_0}_{\alpha_0}(\lambda)| \ \leq \ A\lambda$. This technical condition is interesting because of the following theorem [BJR].

Theorem 1: *Let (X, β, μ, T) denote a dynamical system, with set X, a σ-algebra of its subsets β, a measure μ defined on the measurable space (X, β) such that $\mu(X) \ = \ 1$ and a measurable, measure preserving map $T : X \to X$. Suppose f is in $L^1(X, \beta, \mu)$ and that the sequence of pairs on natural numbers $(n_l, k_l)_{l=1}^\infty$ is Stoltz then*

$$m_f(x) = \lim_{l \to \infty} \frac{1}{k_l} \sum_{i=1}^{k_l} f(T^{n_l + i} x),$$

exists almost everywhere with respect to μ.

See [KN1],[KN2] and [KN3] for applications of this theorem to the metric theory of continued fractions. Note that if $m_{l,f}(x) = \frac{1}{k_l} \sum_{i=1}^{k_l} f(T^{n_l+i}x)$ then

$$m_{l,f}(Tx) - m_{l,f}(x) = k_l^{-1}(f(T^{n_l+k_l+1}x) - f(T^{n_l+1}x)).$$

This means that $m_f(Tx) = m_f(x)$ μ almost everywhere. A dynamical system (X, β, μ, T) is called ergodic if given any $A \in \beta$ we have $T^{-1}A := \{x \in X : Tx \in A\} = A$, the set A has either full or null measure. A standard fact in ergodic theory is that if (X, β, μ, T) is ergodic and for μ measureable k on X we have $k(Tx) = k(x)$ almost everywhere, then $k(x) = \int_X k d\mu$ μ almost everywhere [CFS]. Averages where $n_l \ = \ 1$ for all l will be called non-moving. Moving averages satisfying the above hypothesis can be constructed by taking

for instance $n_l = 2^{2^l}$ and $k_l = 2^{2^{l-1}}$. Proving pointwise ergodic theorems like Theorem 1 is closely related to the proof of results in differentiation theory. For instance the proof of Birkhoff's pointwise ergodic theorem is effected using a maximal inequality called the maximal ergodic theorem. An idea due to N. Wiener clarified by A. Calderon [C] reduces the proving the maximal ergodic theorem to proving it for the special case where $X = \mathbf{Z}$ and T is addition by 1, that is for $x \in \mathbf{Z}$ we have $Tx = x + 1$. In this setting the maximal ergodic theorem is nothing other than than the Hardy Littlewood maximal inequality on the group $\mathbf{Z}$. The proof of this is essentially identical to the case of the more familiar case where the group is the $\mathbf{R}$. As is well known, given a continuous function on the unit circle, it can be realised as the boundry value funtion of a harmonic funtion defined inside the unit disk. The limit behaviour of this harmonic function as the argument inside the unit disc tends toward the unit circle has long been of interest to analysts. It turns out the almost everywhere convergence behavour of this limit is also governed by the Hardy Littlewood maximal finctions. To prove this theorem, it is required that this argument approaches the unit circle confined to a cone within the unit circle. This cone is called the Stoltz cone. It is this condition in Harmonic analysis that inspires the authors of [BJR] to prove Theorem 1. See [NS] for more background.

Prior to the proof of Theorem 1 a number of authors considered moving averages for specific sequences $(n_l, k_l)_{l \geq 1}$. For instance in [AdJ] it is shown that if $n_l = l$ and $k_l = \sqrt{l}$ then there is an L^∞ function for which pointwise convergence of moving averages fails. In [Sc] it is shown that if $n_l = p(l)$ for a non-constant polynomial with coefficients in $\mathbf{Z}$ and $\frac{k_l}{n_l}$ tends to 0 as l tends to infinity, then there is an L^∞ function for which convergence fails. Another L^∞ counter example appears [BV] in the case $n_l = 4^l$ and $k_l = 2^l$. On the other hand as the authors of [BJR] state it was known at the time of writing of [BJR] that if $n_l = 2^{2^l}$ and $k_l = \sqrt{n}_k$ then pointwise convergence takes place. All this is now resolved in Theorem 1.

Let $\cdots x_{-1}, x_0, x_1, \cdots$ be two sided stationary process taking values from the finite set $K = \{a_1, \cdots, a_s\}$ and let $p(x_0, \cdots, x_n)$ denote the joint distribution function of the variables $x_0, \cdots, x_n$. In this paper we prove the following theorem.

Theorem 2 : *Suppose $(n_l, k_l)_{l=1}^\infty$ is Stoltz. Then there is a constant H such that*

$$\lim_{l \to \infty} \frac{1}{k_l} \log p(x_{n_l}, \cdots, x_{n_l + k_l}) = -H,$$

almost everywhere.

In the case $n_l = 1$ for all l, Theorem 2 reduces to the famous Shannon-McMillan-Breiman theorem, referred to briefly as the SMB theorem [Sh][M][B] is the fundamental theorem of information theory. A primary application of the Shannon-McMillan-Breiman theorem is to give a theoretical underpinning to binary data compression of an ergodic time series of entropy $H > 0$. Because the SMB theorem describes generic behaviour one is lead to the concept of a typical set. In particular if $x_1, \ldots, x_n$ is a sequence of stationary variables taking values in a finite state space K. Given $\epsilon \in (0, 1)$ then a typical set with respect to the probability p is the set

$$A_\epsilon^n := \{(x_1, \ldots, x_n) \in K^n : 2^{-n(H+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H-\epsilon)}\}.$$

Elementary arguments, to be found in standard textbooks – see for instance [CT], enable one to conclude that

$$P(A_\epsilon^n) \geq 1 - \epsilon, \tag{1}$$

$$|A_\epsilon^n| \leq 2^{n(H+\epsilon)}; \tag{2}$$

and

$$|A_\epsilon^n| \geq (1 - \epsilon)2^{n(H-\epsilon)} \tag{3}$$

all for large n. These inequalities can be used to describe how to faithfully compress the data from this sequence $x_1, \ldots x_n$ using a binary code. See [CT] for instance for details how this can be done. The role of the SMB theorem here is to ensure the existance of typical sets described above used in the compression. We can now consider an alternative scenario where we have a stationary series of random variables $(x_n)_{n\geq 1}$ which we are only able to observe from time to time. Say for instance a space ship travels towards a far off data source. This data is then collected compressed and returned to earth. Suppose data becomes more plentyful as it approaches the source but that to conserve resources the data is collected only intermittently. A protocol is needed to manage the collection, compression and communication of this data. Theorem 2 tells us how this might be done. Suppose $S = (k_l, n_l)_{l\geq 1}$ denotes a sequence of Stoltz intervals and that data collection and communication are switched off outside Stoltz intervals. We can, without loss of generality assume that the Stoltz intervals are disjoint. Assocated to these Stoltz intervals we can define a typical set

$$B_{S,\epsilon}^l := \{(x_{n_l+1}, \ldots, x_{n_l+k_l}) \in K^{k_l} : 2^{-k_l(H+\epsilon)} \leq p(x_{n_l+1}, \ldots, x_{n_l+k_l}) \leq 2^{-k_l(H-\epsilon)}\}.$$

In light of Theorem 2, it is possible to prove along similar lines, analogues of inequalites (1), (2) and (3), for the sets $(B^l_{S,\epsilon})_{l\geq 1}$ which can then be used to construct a compression scheme along the lines of the one that is constructed from $(A^n_\epsilon)_{n\geq 1}$.

§**2 Proof of Theorem 2 :** In the case of one sided shifts we can think of it as the future of a the two sided shift arising form the natural extention of the one sided shift. We will denote by $\mathbf{E}(f|\mathcal{A})(x)$ the conditional expectation operator of the function f with respect to the σ- algebra $\mathcal{A}$. To prove Theorem 2 we need the following lemma

Lemma 3 : *Suppose* $(\Omega, \mathcal{B}, p, T)$ *is a dynamical system and that* $(g_k)_{k=1}^\infty$ *is a sequence of* p *measurable functions converging pointwise to* g*. Then if* $\sup_{k\geq 1}|g_k| \in L^1(\Omega, \mathcal{B}, p)$ *we have*

$$\lim_{l\to\infty} \frac{1}{k_l} \sum_{k=n_l}^{n_l+k_l} g_k(T^k\omega) = \mathbf{E}(g|\mathcal{I})(\omega).$$

Proof : We have

$$\frac{1}{k_l} \sum_{k=n_l}^{k=n_l+k_l} g_k(T^k x) = \frac{1}{k_l} \sum_{k=n_l}^{k=n_l+k_l} g(T^k x) + \frac{1}{k_l} \sum_{k=n_l}^{k=n_l+k_l} [g_k(T^k x) - g(T^k x)].$$

Using the moving average ergodic theorem the first term on the right tends to $\mathbf{E}(g|\mathcal{I})(x)$. Let $G_N(x) = \sup_{k\geq N}|g_k(x) = g(x)|$. Then for the second term on the right we have the estimate

$$\limsup_{l\to\infty} |\frac{1}{k_l} \sum_{k=n_l}^{k=n_l+k_l} [g_k(T^k x) - g(T^k x)]|$$

$$\leq \limsup_{l\to\infty} |\frac{1}{k_l} \sum_{k=n_l}^{k=n_l+k_l} |g_k(T^k x) - g(T^k x)|$$

$$\leq \limsup_{l\to\infty} |\frac{1}{k_l} \sum_{k=n_l}^{k=n_l+k_l} |G_N(T^k x)| = \mathbf{E}(G_N|\mathcal{I})(x)$$

almost everywhere. Now $(G_N)_{N\geq 1}$ converges monotonically to zero and

$$\mathbf{E}G_0 \leq \mathbf{E}(\sup_k |g_k| + |g|)(x) < \infty,$$

by the monotone convergence theorem $\mathbf{E}(G_N|\mathcal{I})(x)$ converges to 0. Lemma 3 is proved.

We now complete the proof of Theorem 2. Set $g_0(x) = -\log p(x_0)$ and set $g_k(x) = \log \frac{p(x_{-k},\cdots,x_0)}{p(x_{-k}\cdots,x_1)}$ $(k \geq 1)$, where if $(x_n)_{n=0}^{\infty}$ is a one sided sequence we work with the two sided sequences obtained via the natural extention T of the shift map.

$$-\frac{1}{k_l}\log p(x_{n_l},\cdots,x_{n_l+k_l}) = -\frac{1}{k_l}T^{n_l}\log p(x_0,\cdots,x_{k_l-1}) = \frac{1}{k_l}T^{k_l}(\sum_{k=0}^{k_l-1} g_k(T^k x)).$$

Since T is $1-1$ and measure preserving, the proof of Theorem 2 is completed, once we show $(g_k)_{k\geq 0}$ converges almost everywhere an that $\mathbf{E}(\sup_k g_k) < \infty$. To do this we start with an equality of McMillan [M].

$$\int_{m\leq g_k<m+1} g_k \leq s(m+1)2^{-m}.$$

We confine attention to the cylinder set $Z_i \subseteq \Omega$ with $Z_i = \{x : x_0 = a_i\}$. On Z_i we have

$$g_k(x) = -\log p(x_0 = a_i | x_{-1},\cdots,x_k). \tag{4}$$

As $p(x_0 = a_i|x_{-1},\cdots,x_k)_{k\geq 1}$ is a martingale, and $-\log$ is a convex function, the sequence $(g_k(x))_{k\geq 1}$ is a semi-martingle. Then $(g_k)_{k\geq 1}$ converges almost everywhere on Z_i and hence Ω [D]. Furthermore by the semi-martingale property

$$\int_{Z_i} \sup_{0\leq k\leq n} g_k \leq \frac{e}{e-1} + \frac{e}{e-1}\int_{Z_i}(g_n(\log^+ g_n)).$$

Using (4) again we have

$$\int_{Z_i}(g_n \log^+ g_n) = \sum_{m=0}^{\infty}\int_{Z_i\cap[m\leq g_n<m+1]}(g_n\log^+ g_n)$$

$$\leq \sum_{m=0}^{\infty} s(m+1)\log(m+1)2^{-m}.$$

Thus $\int_{Z_i}(\sum_k g_k) < \infty$ so by addition $\mathbf{E}(\sup_k g_k) < \infty$ and Theorem 2 is proved.

REFERENCES

[AdJ]M. A. Akcoglu and A.del Junco: *"Convergence of averages of point transformations"*, Proc. Amer. Math. Soc. **49** (1975), 265-266.

[BJR] A. Bellow, R. Jones and J. Rosenblatt: *"Convergence of moving averages"* Erg. Th. & Dynam. Syst. **10** (1990) no. 1 43–62.

[BL] A. Bellow and V. Losert : *On sequences of zero density on ergodic theory"*, Contemp. Maths. **28** (1984) 49–60.

[Br] L. Breiman : *"The individual ergodic theorem of information theory "*, Ann. Math. Stats. Vol. **28**, no. 3 (1957), 809-811.

[C] A. P. Calderón : *"Ergodic theory and translation-invariant operators"*, Proc. Nat. Acad. Sci. U.S.A. **59** 1968 349-353.

[CFS] I. P. Cornfeld, S. V. Fomin, Ya. G. Sinai, "Ergodic theory", Grundlehren der Mathematischen Wissenschaften, **245**. Springer-Verlag, New York, 1982. x+486 pp.

[CT] T. Cover and J. Thomas, "Elements of Information Theory", Second Edition, Wiley (2006).

[D] J.L. Doob : *"Stochastic Processes"*, John Wiley & Sons Inc. New York.

[KN1] H. Kamarul-Haili and R. Nair, : *On moving averages and continued fractions* Unif. Distrib. Theory **6** (2011), no. 1, 65-78..

[KN2] H. Kamarul-Haili and R. Nair : *Optimal continued fractions and the moving average ergodic theorem*, Period. Math. Hungar. **66** (2013), no. 1, 95-103.

[KN3] H. Kamarul-Haili and R. Nair : *The nearest integer continued fraction and the moving average ergodic theorem*, Unif. Distrib. Theory **8** (2013), no. 1, 73-87.

[M] B. McMillan : *"The basic theorems of Information "*, Ann. Math. Stats. Vol. **24** (1953), 196-219.

[NS] A. Nagel and E.M. Stein : *"On certain maximal functions and approach regions"*, Adv. in Math. **54** (1984), no. 1, 83-106.

[Sc] M. Schwartz : *"Polynomially moving ergodic averages"*, Proc. Amer. Math. Soc. **103** (1988), no. 1, 252-254.

[Sh] C. E. Shannon : *"A mathematical theory of communication"*, Bell Systems Technical Journal Vol. **27** (1948) 379-423 and 623-656.